Εργασία στο μάθημα
Διαχείριση Μεγάλων Δεδομένων

Αναφορά

ΜΑΡΑΓΚΑΚΗΣ ΑΛΕΞΑΝΔΡΟΣ 2022201900120
ΑΡΓΥΡΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ 2022202000014

Περιεχόμενα:

(1) Λεπτομέρειες Υλοποίησης – Οδηγίες

Για την υλοποίηση της εργασίας και την εξαγωγή αποτελεσμάτων, χρησιμοποιήσαμε τις εξής τεχνολογίες:

- python (https://www.python.org/),
- jupyter notebook (https://jupyter.org/),
- docker (https://www.docker.com/),
- mongo compass (https://www.mongodb.com/products/compass/).

Στήσαμε την mongodb μέσα σε ένα docker container, ώστε να απλοποιήσουμε την διαδικάσια εγκατάστασης.

Για την μετατροπή των .csv αρχείων σε .json, ο κώδικας κάνει αποσυμπίεση το αρχείο archive.zip που βρίσκονται τα δεδομένα μας, και με την βοήθεια της βιβλιοθήκης pandas (https://pandas.pydata.org/) κάνει την μετατροπή, και παράλληλα προσθέτει ένα νέο πεδίο με τον κωδικό της χώρας του κάθε αρχείο στα αντίστοιχα έγγραφα.

Έπειτα με την βοήθεια της βιβλιοθήκης pymongo (https://pymongo.readthedocs.io/en/stable/) συνδέουμε τον κώδικα με την mondodb, δημιουργούμε την βάση και τα collections, και φορτώνουμε τα αρχεία. Αξίζει να σημειωθεί ότι η βάση και τα collections δημιουργούνται την στιγμή που ανεβαίνουν τα αρχεία, και όχι στην αρχικοποίηση που γίνεται στον κώδικα.

Το ανέβασμα όλων των αρχείων διαρκεί περίπου 2,5 λεπτά.

Για την επίλυση των ζητημάτων της εργασίας, φτιάξαμε τα aggregations μέσω του Mongo Compass, και έπειτα χρησιμοποιήσαμε την επιλογή "export to language" για την μετατροπή σε python κώδικα. Τα αποτελέσματα των aggregations επιστρέφονταν σε μορφή αντικειμένου iterator, το οποίο μετά την οποιαδήποτε πράξη πάνω στην επιστρεφόμενη τιμή, την διέγραφε.

Για την οπτικοποίηση των δεδομένων χρησιμοποιήσαμε την βιβλιοθήκη matplotlib (https://matplotlib.org/stable/) της python, ενώ για το bonus ερώτημα χρησιμοποιήσαμε την βιβλιοθήκη numpy (https://numpy.org/) για να εξάγουμε περιγραφικά στατιστικά.

Ο κώδικας υπάρχει μέσα σε ενα Jupyter Notebook, και οι απαιτήσεις για να τρέξει υπάρχουν στο αρχείο requirements.txt.

Για να αναπαραχθούν τα ίδια αποτελέσματα, θα πρέπει να κατεβάσετε το αρχείο απο την πηγή του, και να κόψετε τα δεδομένα που προστέθηκαν από τον δημιουργό μετά τις 12/01/23 (dd/mm/yy). Επίσης θα πρέπει να εγκαταστήσετε το Jupyter Notebook, ή κάποιο αντίστοιχο extention στο IDE σας, την python, και τις απαιτήσεις από το αρχείο requirements.txt με την χρήση του pip package manager.

Η εντολή είναι "$pip install -r requirements.txt" , και προτείνουμε δημιουργήσετε πρώτα ένα virtual enviroment με τις εντολές "$python3 -m venv env", "$source env/bin/activate", "$pip install -r requirements.txt", και μετά να εκκινήσετε το jupyter notebook στο ίδιο shell, ή να επιλέξετε στον IDE σας τον python interpreter μέσα στο env.

(2) Ερώτημα 1
Τι ξέρουμε για τις δημοσιεύσεις του πολύ δημοφιλούς καναλιού Saturday Night Live16;

(2.1) Aggregation Stages

$match stage: Φιλτράρει τα έγγραφα της συλλογής με βάση συγκεκριμένα κριτήρια. Σε αυτή την περίπτωση, φιλτράρει έγγραφα όπου το πεδίο ‘country’ είναι ίσο με 'GB' και το πεδίο ‘channelTitle’ είναι ίσο με “Saturday Night Live”.

$sort stage: Σε αυτή την περίπτωση, ταξινομεί τα έγγραφα σε φθίνουσα σειρά με βάση το πεδίο trending_date.

$group stage: Ομαδοποιεί τα έγγραφα με βάση ένα καθορισμένο πεδίο, στην περίπτωση αυτή, ομαδοποιεί τα έγγραφα με βάση τον τίτλο τους. Υπολογίζει επίσης συγκεντρωτικές τιμές για κάθε ομάδα, όπως τα πρώτα πεδία ‘likes’, ‘dislikes’, ‘view_count’ και ‘publishedAt’.

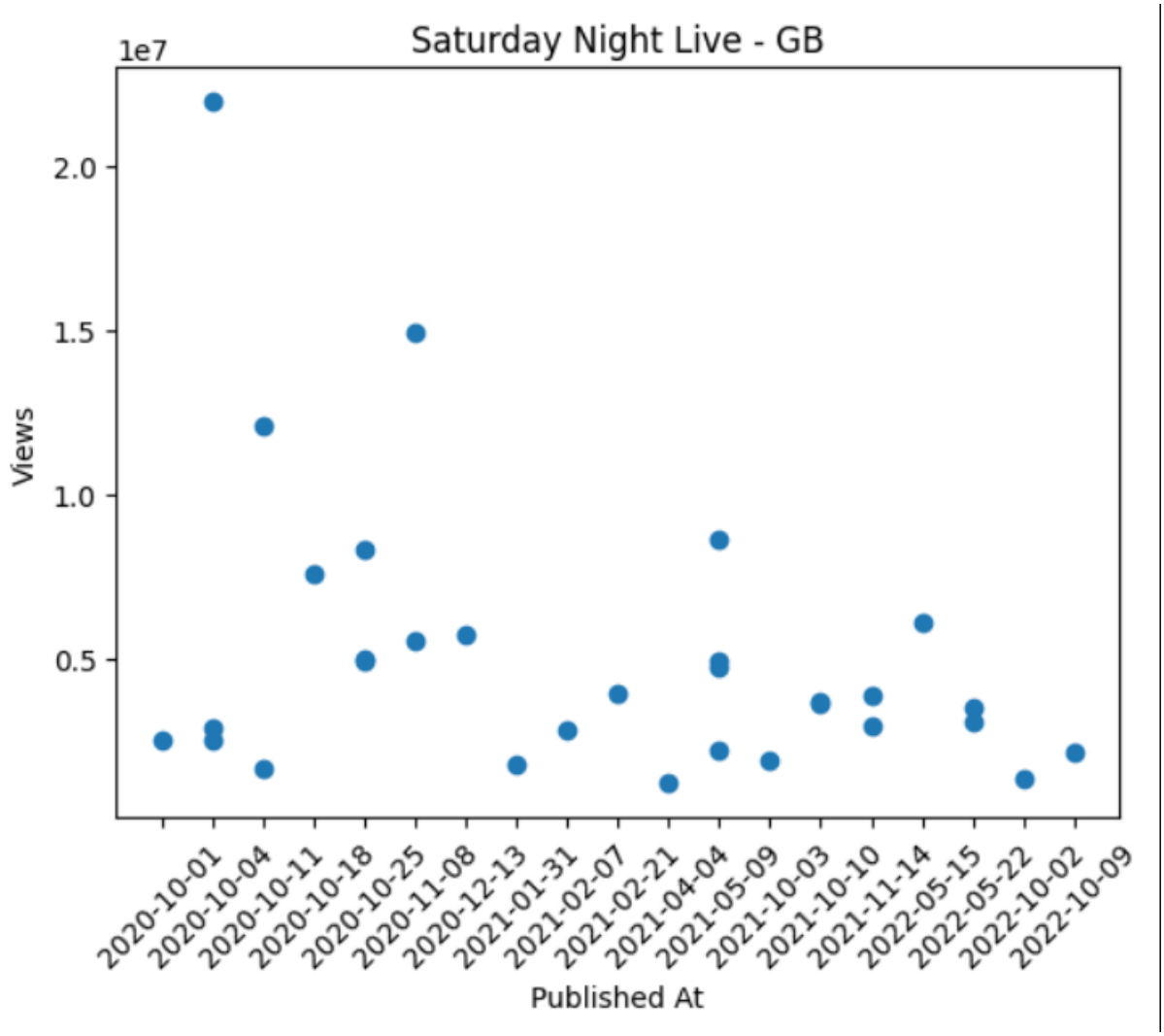
$sort stage: Ταξινομεί τα προκύπτοντα έγγραφα σε αύξουσα σειρά με βάση το πεδίο ‘publishedAt’.

$project stage: Καθορίζει τα πεδία που θα συμπεριληφθούν ή θα αποκλειστούν στην τελική έξοδο του aggregation. Σε αυτή την περίπτωση, περιλαμβάνει τα πεδία ‘_id’, ‘likes’, ‘dislikes’, ‘views’ και ‘publishedAt’. Μετατρέπει επίσης το πεδίο ‘publishedAt’ σε συμβολοσειρά με τη μορφή 'YYYY-MM-DD'.

(2.2) Αποτελέσματα

	_id	likes	dislikes	views	publishedAt
2	Jim Carrey and Maya Rudolph Transform into Joe Biden and Kamala Harris - SNL	42891	2650	2526816	2020-10-01
3	Superspreader Event - SNL	43426	1370	2504202	2020-10-04
4	Weekend Update: Trump Tests Positive for Covid - SNL	40076	2449	2916986	2020-10-04
5	First Debate Cold Open - SNL	437775	57125	21954683	2020-10-04
6	New Normal - SNL	29237	1326	1646765	2020-10-11
7	VP Fly Debate Cold Open - SNL	188194	31682	12070479	2020-10-11
8	Dueling Town Halls Cold Open - SNL	100317	11249	7610686	2020-10-18
9	Adele Monologue - SNL	65916	1621	5002750	2020-10-25
10	Final Debate Cold Open - SNL	136664	20003	8321084	2020-10-25
11	The Bachelor - SNL	93511	1247	4929699	2020-10-25
12	Biden Victory Cold Open - SNL	275185	29394	14927666	2020-11-08
13	Weekend Update: Rudy Giuliani on Trump's Election Lawsuits - SNL	83187	4679	5530282	2020-11-08
14	Rap Roundtable - SNL	214936	4382	5727960	2020-12-13
15	Twins - SNL	38236	870	1794947	2021-01-31
16	Super Bowl Pre-game Show Cold Open - SNL	35177	2714	2845426	2021-02-07
17	Britney Spears Cold Open - SNL	53080	5830	3912795	2021-02-21
18	Salt Bae - SNL	22567	1144	1207224	2021-04-04
19	Elon Musk Monologue - SNL	195326	16510	8649926	2021-05-09
20	Weekend Update: Financial Expert Lloyd Ostertag on Cryptocurrency - SNL	47815	2361	2193521	2021-05-09
21	Chad on Mars - SNL	154928	5103	4765679	2021-05-09

(2.3) Γραφική Απεικόνιση Αποτελεσμάτων



(2.4) Τελικά Συμπεράσματα

Εξετάζοντας το παραπάνω διάγραμμα, παρατηρούμε ότι τα βίντεο με τις περισσότερες προβολές δημοσιεύτηκαν κατά τη περίοδο της καραντίνας. Μετά τη περίδο της καραντίνας οι προβολές φαίνονται να μειώνονται. Αυτό πιθανόν να εξηγεί και την θετική κατανομή των δεδομένων στο γράφημα.

(3) Ερώτημα 2
Πόσες ετικέτες χρησιμοποιούνται συνήθως στις δημοσιεύσεις των βίντεο;

(3.1) Aggregation Stages

$match stage: Φιλτράρει τα έγγραφα της συλλογής με βάση συγκεκριμένα κριτήρια. Σε αυτή την περίπτωση, φιλτράρει έγγραφα όπου το πεδίο ‘country’ είναι ίσο με 'GB', το πεδίο ‘tags’ δεν είναι ίσο με τη λίστα με τη μοναδική τιμή '[None]' και το πεδίο ‘view_count’ είναι μεγαλύτερο από 0. Όταν το view_count είναι ίσο με το 0, πρόκειται για ζωντανό (live) βίντεο.

$sort stage: Ταξινομεί τα έγγραφα της συλλογής με βάση ένα συγκεκριμένο πεδίο, σε αυτή την περίπτωση, ταξινομεί τα έγγραφα σε φθίνουσα σειρά με βάση το πεδίο ‘trending_date’.

$group stage: Ομαδοποιεί τα έγγραφα με βάση ένα καθορισμένο πεδίο, στην περίπτωση αυτή, ομαδοποιεί τα έγγραφα με βάση τον τίτλο τους. Υπολογίζει επίσης συγκεντρωτικές τιμές για κάθε ομάδα, όπως ο μέσος αριθμός ετικετών στο πεδίο ‘tags’, την πρώτη τιμή του πεδίου ‘view_count’ και την πρώτη τιμή του πεδίου ‘video_id’.

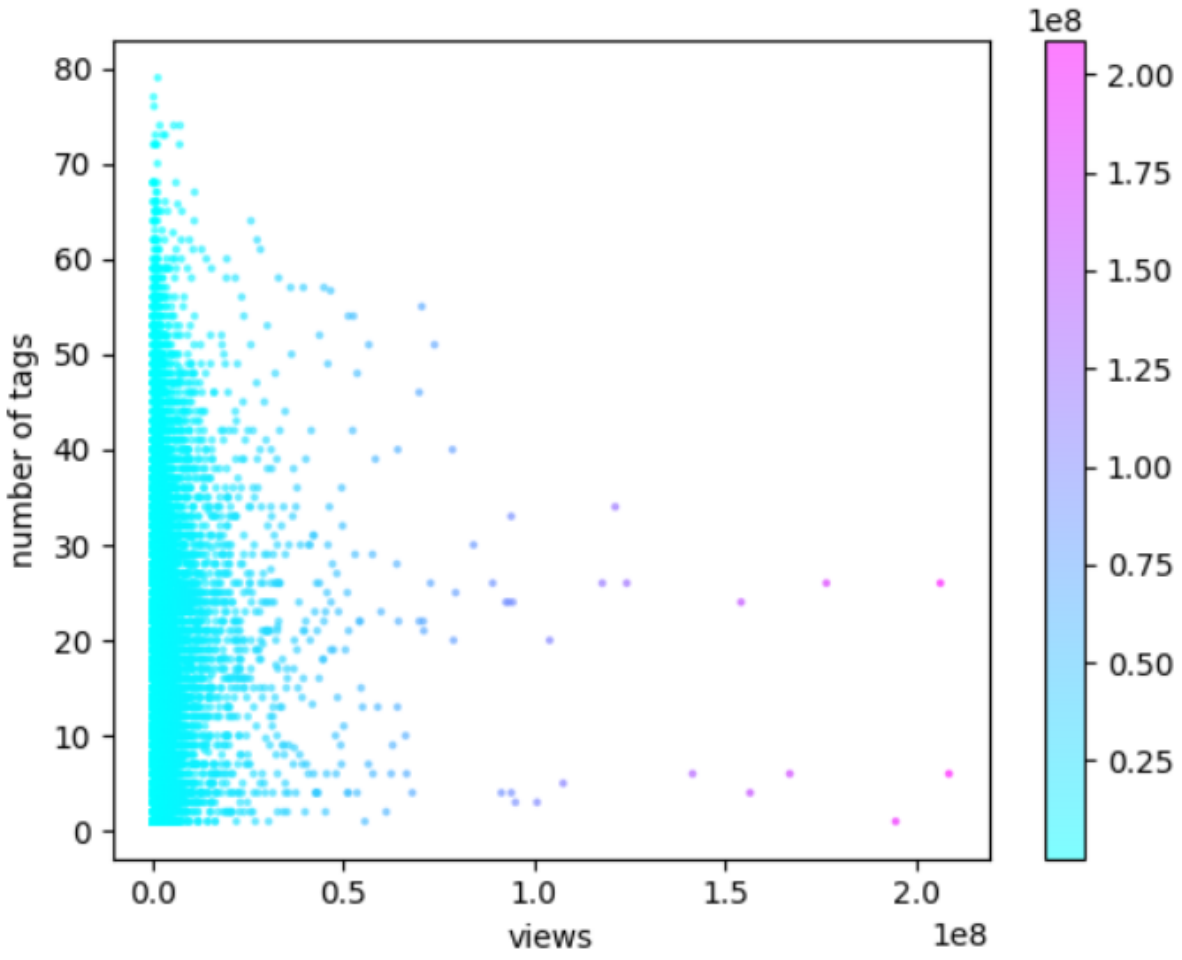
$sort stage: Ταξινομεί τα προκύπτοντα έγγραφα σε φθίνουσα σειρά με βάση το πεδίο ‘views’.

$project stage: Καθορίζει τα πεδία που θα συμπεριληφθούν ή θα αποκλειστούν στην τελική έξοδο του aggregation. Σε αυτή την περίπτωση, αποκλείει το πεδίο ‘_id’ και περιλαμβάνει τα πεδία ‘avg_num_tags’, ‘video_id’ και ‘views’.

(3.2) Αποτελέσματα

```
avg_num_tags,views,video_id
6,208581468,gdZLi9oWNZg
26,206344829,gQlMMD8auMs
1,194625542,qF0N19MgI3Q
26,176467113,vRXZj0DzXIA
6,166895681,WMweEpGlu_U
4,156482499,CuklIb9d3fI
24,154134590,awkkyBH2zEo
6,141428767,-5q5mZbe3V8
26,124180499,dyRsYk0LyA8
34,121159003,8dJyRm2jJ-U
26,117791538,POe9SOEKotk
5,107534237,T6n7lZirQkA
20,104017073,U3ASj1L6_sY
3,100694053,46SbB0IHplE
3,95023322,pFtsvQuefXQ
24,94356729,CKZvWhCqx1s
4,93952431,kXpOEzNZ8hQ
33,93935173,hsm4poTWjMs
24,93279276,P0ya8iGw2QE
24,92559180,myjEoDypUD8
```

(3.3) Γραφική Απεικόνιση Αποτελεσμάτων



(3.4) Τελικά Συμπεράσματα

Εξετάζοντας το διάγραμμα, συμπεραίνουμε αρχικά ότι ο αριθμός των ετικετών δεν επηρεάζει άμεσα τον αριθμό των προβολών. Η αύξηση του αριθμού των ετικετών σε ένα βίντεο, δε συνεπάγεται την αύξηση των προβολών του. Τα βίντεο με τη μεγαλύτερη διαφορά στις προβολές κατέχουν γύρω στα 5-25 tags.

(4) Ερώτημα 3
Πώς φέρονται οι vloggers και οι χρήστες ανά περιοχή;

(4.1) Aggregation Stages

$match stage: Φιλτράρει τα έγγραφα της συλλογής με βάση συγκεκριμένα κριτήρια. Σε αυτή την περίπτωση, φιλτράρει έγγραφα στα οποία το πεδίο tags δεν είναι ίσο με τη λίστα με τη μοναδική τιμή '[None]'.

$group stage: Ομαδοποιεί τα έγγραφα με βάση ένα καθορισμένο σύνολο πεδίων, σε αυτή την περίπτωση, ομαδοποιεί τα έγγραφα με βάση το video_id και τη χώρα τους. Υπολογίζει επίσης συγκεντρωτικές τιμές για κάθε ομάδα, όπως έναν πίνακα ετικετών και το μέγιστο view_count για κάθε ομάδα.

$unwind stage: Αποδομεί τον πίνακα 'num_tags' από το προηγούμενο στάδιο, δημιουργώντας ένα νέο έγγραφο για κάθε στοιχείο του πίνακα.

$unwind stage: Αποδομεί ξανά τον πίνακα 'num_tags', δημιουργώντας ένα νέο έγγραφο για κάθε στοιχείο του πίνακα. Αυτό είναι απαραίτητο επειδή ο πίνακας 'num_tags' περιέχει πίνακες ετικετών.

$group stage: Ομαδοποιεί τα έγγραφα με βάση το πεδίο '_id' από το προηγούμενο στάδιο, το οποίο αποτελείται από το 'video_id' και τη χώρα. Δημιουργεί επίσης ένα νέο πίνακα 'num_tags' που περιέχει όλες τις μοναδικές ετικέτες για κάθε βίντεο και διατηρεί την πρώτη τιμή 'view_count' για κάθε ομάδα.

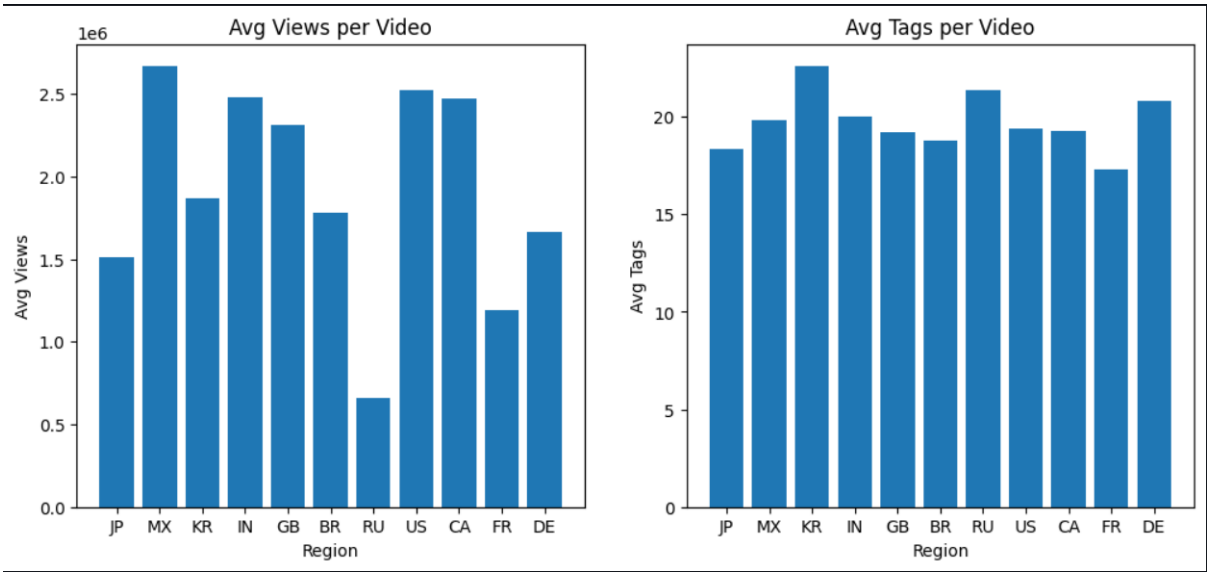
$group stage: Ομαδοποιεί τα έγγραφα με βάση το πεδίο 'country' από το προηγούμενο stage. Υπολογίζει τον μέσο αριθμό ετικετών ανά βίντεο και τον μέσο όρο προβολών ανά βίντεο για κάθε χώρα.

$project stage: Καθορίζει τα πεδία που θα συμπεριληφθούν ή θα αποκλειστούν στην τελική έξοδο του aggregation. Σε αυτή την περίπτωση, περιλαμβάνει τα πεδία 'avg_tags_per_video' και 'avg_views_per_video', τα οποία υπολογίζονται στο προηγούμενο στάδιο, και τα στρογγυλοποιεί σε δύο δεκαδικά ψηφία χρησιμοποιώντας τον τελεστή $round.

(4.2) Αποτελέσματα

_id	avg_tags_per_video	avg_views_per_video
RU	21.38	661357.84
BR	18.76	1781443.07
US	19.38	2522774.1
MX	19.83	2668388.03
FR	17.31	1188564.42
CA	19.3	2472857.13
JP	18.36	1514489.45
DE	20.8	1665047.7
KR	22.6	1871149.04
IN	20.02	2479827.79
GB	19.24	2314795.51

(4.3) Γραφική Απεικόνιση Αποτελεσμάτων



(4.4) Τελικά Συμπεράσματα

Εξετάζοντας τα διαγράμματα, παρατηρούμε ότι ο μέσος αριθμός των tags είναι περίπου ίσος ανά περιοχή. Ο μέσος αριθμός των ετικετών ανά περιοχή δεν επηρεάζει στον ίδιο βαθμό τον μέσο αριθμό προβολών ανά περιοχή. Αυτό φαίνεται, για παράδειγμα, από την Κορέα που παρόλο που είναι πρώτη στο μέσο όρο των ετικετών, βρίσκεται έκτη στον μέσο όρο προβολών. Αξίζει να σημειωθεί ότι η Ρωσία αποτελεί εξαίρεση μιας και λόγο του πρόσφατου πολέμου με την Ουκρανία, δεν έχει τόση συμμετοχή στα κοινωνικά δίκτυα όπως το YouTube.

(5) Ερώτημα 4

Ποιες είναι οι πιο δημοφιλείς ετικέτες στα ανερχόμενα βίντεο;

(5.1) Aggregation Stages

$match stage: Φιλτράρει τα έγγραφα με βάση το πεδίο ‘country’ για να συμπεριλάβει μόνο τα έγγραφα στα οποία η χώρα είναι είτε 'US' είτε 'GB'.

$group stage: Ομαδοποιεί τα έγγραφα με βάση το πεδίο ‘country’ και δημιουργεί έναν πίνακα όλων των ετικετών για κάθε χώρα.

$unwind stage: Αποδομεί τον πίνακα ετικετών, έτσι ώστε κάθε ετικέτα να έχει το δικό της έγγραφο.

$unwind stage: Αποδομεί ξανά τον πίνακα ‘tags’ για να χωρίσει κάθε ετικέτα σε ξεχωριστό έγγραφο. Αυτό είναι απαραίτητο επειδή ο πίνακας ‘tags’ περιέχει πίνακες ετικετών.

$group stage: Ομαδοποιεί τα έγγραφα με βάση τα πεδία ‘tags’ και ‘country’ και μετρά τον αριθμό των εμφανίσεων κάθε ετικέτας για κάθε χώρα.

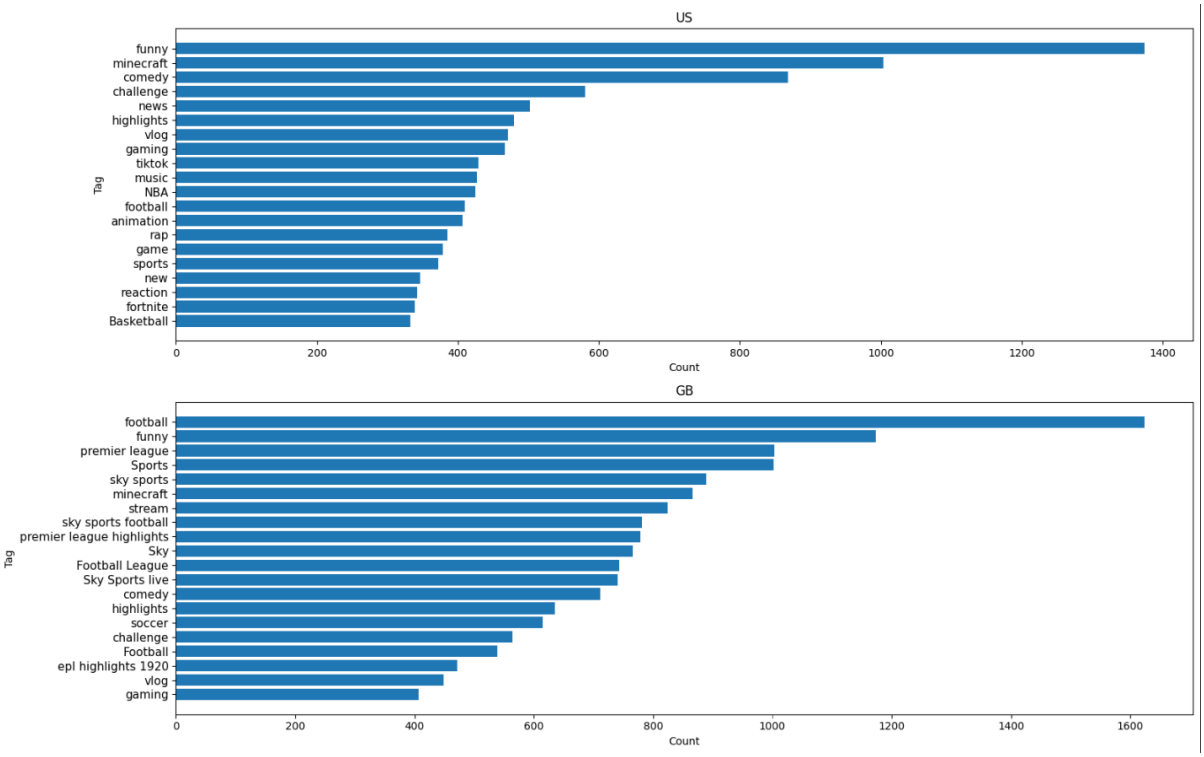
$sort stage: Ταξινομεί τα έγγραφα σε φθίνουσα σειρά με βάση το πεδίο ‘count’.

$project stage: Καθορίζει ποια πεδία θα συμπεριληφθούν στο τελικό αποτέλεσμα και μετονομάζει το πεδίο ‘_id’ σε ‘tag’ και ‘country’ για λόγους ευανάγνωστης χρήσης.

(5.2) Αποτελέσματα

```
count,country,tag
1624,GB,football
1374,US,funny
1174,GB,funny
1004,GB,premier league
1003,US,minecraft
1002,GB,Sports
889,GB,sky sports
868,US,comedy
866,GB,minecraft
825,GB,stream
781,GB,sky sports football
779,GB,premier league highlig
766,GB,Sky
743,GB,Football League
741,GB,Sky Sports live
711,GB,comedy
635,GB,highlights
615,GB,soccer
580,US,challenge
564,GB,challenge
539,GB,Football
502,US,news
480,US,highlights
472,GB,epl highlights 1920
471,US,vlog
467,US,gaming
449,GB,vlog
429,US,tiktok
427,US,music
425,US,NBA
410,US,football
407,GB,gaming
406,US,animation
385,US,rap
379,US,game
377,GB,Manchester United
377,GB,news
372,US,sports
351,GB,music
351,GB,tiktok
346,US,new
342,US,reaction
339,GB,Premier League
339,US,fortnite
334,GB,boxing
```

(5.3) Γραφική Απεικόνιση Αποτελεσμάτων



(5.4) Τελικά Συμπεράσματα

Η πλειοψηφία των ετικετών στην Αμερική αφορά θέματα ψυχαγωγίας που σχετίζονται περισσότερο με το gaming, μουσική και αθλητισμό. Από την άλλη, στην Αγγλία η συντριπτική πλειοψηφία των ετικετών αφορά το ποδόσφαιρο.

(6) Ερώτημα 5

Τι αντίκτυπο έχει στο κοινό η απενεργοποίηση των σχολίων;

(6.1) Aggregation Stages

$match stage: Το πεδίο ‘comments_disabled’ έχει οριστεί σε True/False.

$group stage: Αυτό το στάδιο ομαδοποιεί τα βίντεο με βάση το None (δηλ. ομαδοποιεί όλα τα βίντεο μαζί) και υπολογίζει τον μέσο αριθμό των views, likes και dislikes για την ομάδα.

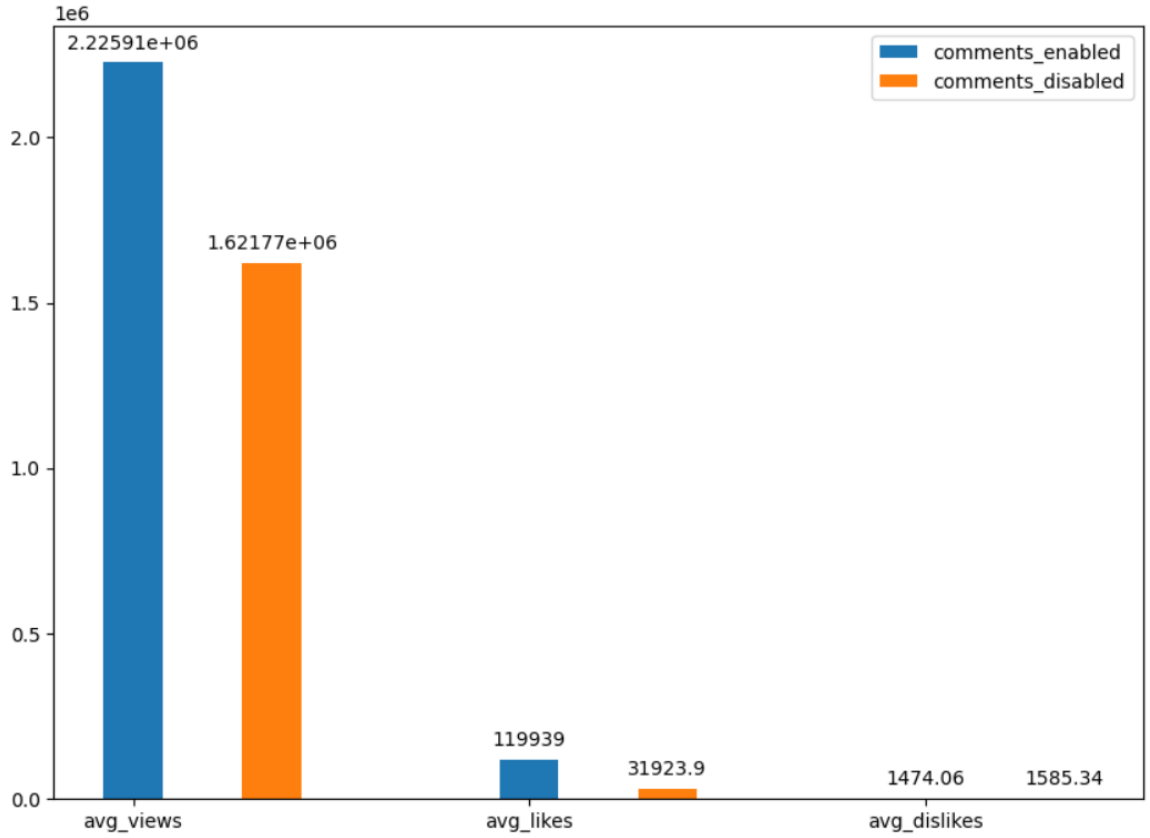
$project stage: Αυτό το στάδιο προβάλλει τον μέσο αριθμό views, likes και dislikes για την ομάδα, περικόπτοντας το καθένα σε 2 δεκαδικά ψηφία.

(6.2) Αποτελέσματα

	avg_views	avg_likes	avg_dislikes
2	1621772.54	31923.86	1585.34

	avg_views	avg_likes	avg_dislikes
2	2225908.12	119939.32	1474.06

(6.3) Γραφική Απεικόνιση Αποτελεσμάτων



(6.4) Τελικά Συμπεράσματα

Στο διάγραμμα παρατηρούμε ότι η διαφορά ανάμεσα στα βίντεο με ενεργοποιημένα σχόλια και στα βίντεο με απενεργοποιημένα σχόλια είναι μεγάλη, όσον αφορά τις προβολές και τα likes. Ωστόσο για τα dislikes, αν λάβουμε υπόψη τις μεγάλες διαφορές από τις προηγούμενες δύο κατηγορίες, τότε συμπεραίνουμε ότι τα dislikes έχουν μεγαλύτερο αντίκτυπο στα βίντεο με απενεργοποιημένα σχόλια παρά τη μικρή διαφορά τους με την άλλη κατηγορία βίντεο.

(7) Ερώτημα 6

Ποιες ήταν οι πιο δημοφιλείς ημερομηνίες για δημοσίευση βίντεο;

(7.1) Aggregation Stages

$matchstage: Φιλτράρισμα των εγγράφων όπου η χώρα είναι 'GB'.

$project stage: Δημιουργία ενός νέου πεδίου με όνομα ‘date’, το οποίο είναι μια μορφοποιημένη συμβολοσειρά του πεδίου ‘publishedAt’ ως συμβολοσειρά ημερομηνίας.

$match stage: Φιλτράρισμα των εγγράφων όπου το πεδίο ‘date’ είναι στις ή μετά τις '2022-09-01'.

$group stage: Ομαδοποίηση των εγγράφων με βάση το πεδίο ‘date’ και μετράει τον αριθμό των εγγράφων σε κάθε ομάδα.

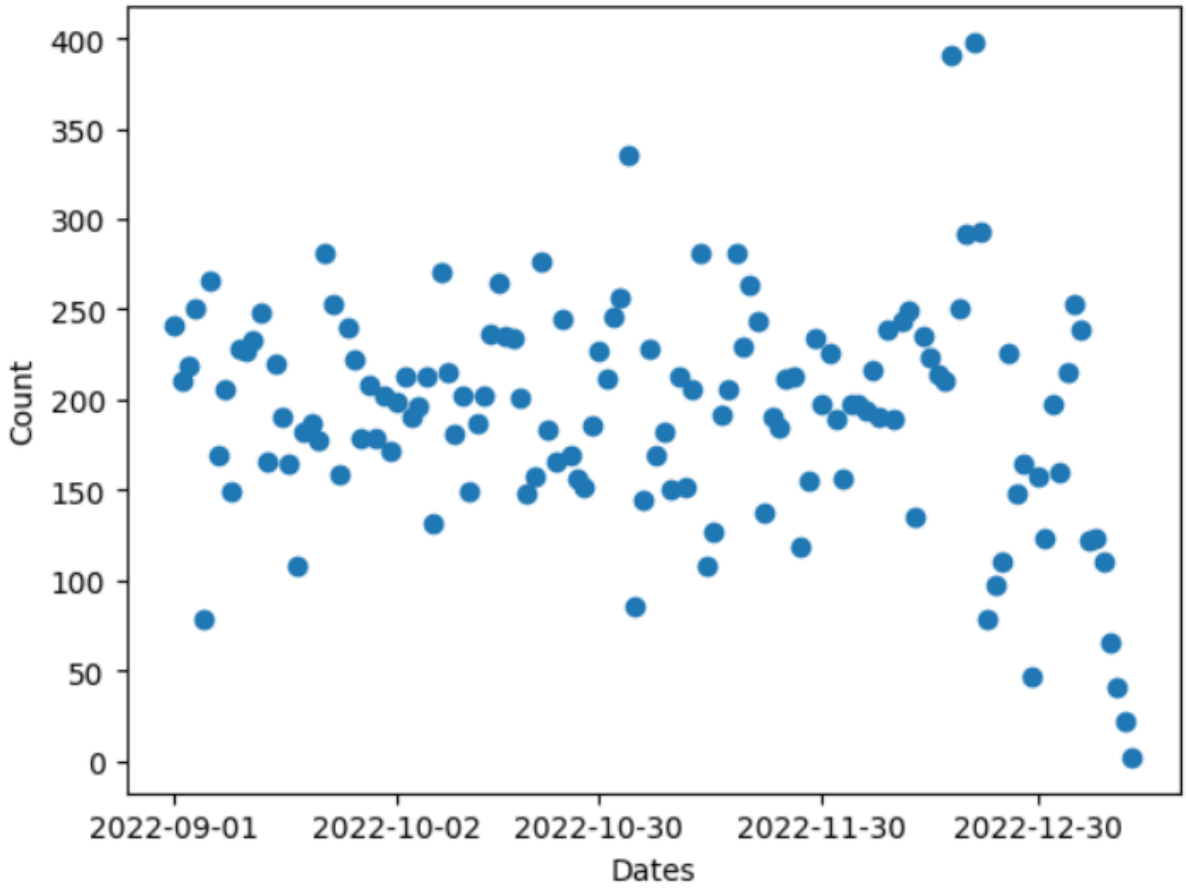
$sort stage: Ταξινόμηση των εγγράφων με βάση το πεδίο ‘_id’ σε αύξουσα σειρά.

$project stage: Μετονομασία του πεδίου ‘_id’ σε date.

(7.2) Αποτελέσματα

count	date
241	2022-09-01
210	2022-09-02
219	2022-09-03
251	2022-09-04
79	2022-09-05
266	2022-09-06
169	2022-09-07
206	2022-09-08
149	2022-09-09
228	2022-09-10
227	2022-09-11
233	2022-09-12
248	2022-09-13
166	2022-09-14
220	2022-09-15
190	2022-09-16
165	2022-09-17
108	2022-09-18
182	2022-09-19
187	2022-09-20
177	2022-09-21

(7.3) Γραφική Απεικόνιση Αποτελεσμάτων



(7.4) Τελικά Συμπεράσματα

Παρατηρώντας τα αποτελέσματα, συμπεραίνουμε ότι το 10ήμερο πριν τα Χριστούγεννα έχουμε τις περισσότερες δημοσιεύσεις ανά ημέρα. Στο τριήμερο όμως των Χριστουγέννων, ο αριθμός των δημοσιεύσεων πέφτει κατακόρυφα.

(8) Ερώτημα 7 (Bonus)
Ημέρες για εισαγωγή στο trending list.

(8.1) Aggregation Stages

$match stage: Χρησιμοποιείται για το φιλτράρισμα των βίντεο που προέρχονται είτε από τις χώρες "ΗΠΑ" είτε από τη Βρετανία.

$group stage: Ομαδοποιεί τα βίντεο με βάση το ‘video_id’ τους και βρίσκει την ελάχιστη ‘trending_date’ και ‘publishedAt’ για κάθε βίντεο.

$project stage: Μετατρέπουμε τα πεδία ‘publishedAt’ και ‘trending_date’ σε αντικείμενα ‘date’ χρησιμοποιώντας τον τελεστή $convert.

$project stage: Υπολογίζουμε τη διαφορά μεταξύ των ‘trending_date’ και ‘publishedAt’ σε χιλιοστά του δευτερολέπτου χρησιμοποιώντας τον τελεστή $subtract. Στη συνέχεια

διαιρούμε αυτή τη διαφορά με τον αριθμό των χιλιοστών του δευτερολέπτου σε μια ημέρα για να λάβουμε τον αριθμό των ημερών που χρειάστηκε το βίντεο για να γίνει trending και στρογγυλοποιούμε προς τα κάτω χρησιμοποιώντας το $floor.

$project stage: Προβάλλουμε μόνο το πεδίο ‘days_to_trend’ και αφαιρούμε το πεδίο ‘_id’ χρησιμοποιώντας το '_id': 0.

(8.2) Αποτελέσματα

```
Mean:  0.5358409925517127
Median:  0.0
Standard deviation:  1.229557712349709
Max:  30.0
```

(8.3) Τελικά Συμπεράσματα

Υπάρχει κάποιο(ά) βίντεο-εξαίρεση που ήταν εκτός λίστας trending για 30 ημέρες μετά την ημερομηνία δημοσίευσης. Ο μέσος όρος ημερών για την εισαγωγή του κάθε βίντεο στη λίστα με τα trends είναι 0.5, ενώ η τυπική απόκλιση είναι 1.22 μέρες (μία ημέρα και κάτι ώρες). Για τις περιοχές US και GB συμπεραίνουμε ότι εφόσον το μέσο διάστημα days-to-trend είναι μισή μέρα, τα βίντεο αυτά δημοσιεύονται με στόχο να μπουν στην trending λίστα και οι δημιουργοί τους πιθανόν αξιοποιούν τεχνικές που έχουν προκύψει από ανάλυση δεδομένων για να το καταφέρουν.